FAISAL RIAZ

+923450637663 | faisal.r.wattoo@gmail.com | LinkedIn | GitHub | Medium | Celystik Labs

SUMMARY OF QUALIFICATIONS

● 2+ years of experience in Data Engineering providing both on-premise and cloud-based solutions. Hands on experience across multiple Data Platforms including Azure Cloud, AWS Cloud, GCP Cloud and Open-Source Apache technologies.

● Collaborating with cross-functional teams and driving data engineering initiatives.

Tools & Technologies:

● Big-Data Stack (Open-Source): Spark, Hadoop, HDFS, Hive, YARN, Airflow.

● Azure Data Platform: ADLS Gen2, Databricks, Azure Synapse Analytics, Azure Data Factory

● AWS Data Platform: AWS Glue, S3, Athena, DMS, CloudWatch, Lambda, EventBridge, SNS, Redshift.

● GCP Data Platform: BigQuery, Cloud Storage, Pub/Sub, Compute Instances, Cloud Data Fusion, Dataproc, Cloud Composer, Cloud Dataflow

● Programming Languages: SQL, Python, Shell-Scripting, Pyspark

Certifications:

●Microsoft Certified: Azure Data Engineer Associate

●Astronomer Certification : Apache Airflow Fundamentals

●Astronomer Certification : DAG Authoring for Apache Airflow

●IBM Hadoop Data Access - Level 2

●IBM Hadoop Foundations - Level 2

●HackerRank | SQL 4 Star, Python

●Databricks: Generative AI Fundamentals

●Coursera: Machine Learning Specialization

Education:

● Bachelors in Software Engineering (2021 - 2025) from National University of Science and Technology (NUST) Pakistan

PROFESSIONAL EXPERIENCE

Data Engineer at Celystik Labs - Lahore - Remote

June 2022 - Present

Project: Redex - United Kingdom

● Developed a templatized ETL framework for fetching tables from SQL Server. Implemented a migration pipeline leveraging Infrastructure as Code (boto3) to automate the creation of required resources, generate Glue jobs from S3-based scripts Developed ETL templates using Appflow for efficient data transfer

Technology Stack: : Glue, Crawlers, Data Catalog, S3, Athena

Project: Enterprise Global Athletic Wear Company

● Developed Spark streaming pipelines in Databricks for seamless ingestion from Azure Data Lake Storage. Leveraged UDFs (User-Defined Functions) to efficiently process the messages and stored them persistently in ACID-compliant Delta tables.

● Developed an ETL process to optimize resource allocation and freelancer hiring by reading and normalizing nested JSON data from on premise web applications and tabular Excel data from SharePoint sites, persisting it in a highly normalized schema using delta tables

● Designed and implemented Azure Data Factory (ADF) pipelines to orchestrate EL-based data migration from on-premise systems including Oracle databases, mount points onto cloud data lake (ADLS Gen 2)

Technology Stack: : ADLS Gen2, Databricks, Azure Synapse Analytics, Azure Data Factory

Project: Electronic Access Request Management

●Developed ETL pipelines using Spark primarily SparkSQL Datasets to perform transformations on 28+ ingested streams of data. Extracted data from a variety of sources including Hive and HDFS using Hadoop scripts.

● Transformed and unified input data according to a common table schema. Wrote spark-submit jobs to load the transformed data into output tables and wrote bash scripts to deploy the pipelines to production.

● Developed Rule Ingestion Alarm Framework to monitor, categorize and notify using spark datasets and SparkSQL. Filtering logic implemented in spark and parameterized using inputs from external sources such as Hive to filter data exceeding a certain threshold

Technology Stack: HDFS, Spark, Hive, SQL

Project: Cloud-Based Supply Chain Analytics Platform

●Centralized supply chain data from MySQL, APIs, and flat files using Cloud Data Fusion.Automated ETL workflows with Cloud Composer and processed large-scale data transformations on Dataproc.

●Designed and optimized a star schema in BigQuery with partitioning and clustering for analytics.

Technology Stack: BigQuery, Cloud Data Fusion, Dataproc, Cloud Storage, Cloud Composer

Project: Ethereum Dataflow Pipeline

●**Developed** a serverless data pipeline to ingest Ethereum blockchain data into Amazon S3 using AWS Lambda for extraction and storage.

●**Automated** data transformation processes using AWS Glue jobs and monitored workflow status with AWS Step Functions.

●**Integrated** Amazon Athena to query transformed data directly from S3, enabling on-demand analytics and reporting.

●**Implemented** error handling and retry mechanisms for AWS Glue and Athena operations, ensuring data integrity and processing reliability.

Technology Stack: AWS Lambda, Amazon S3, AWS Glue, AWS Step Functions, Amazon Athena